PARAMETER UNCERTAINTY QUANTIFICATION USING
SURROGATE MODELS APPLIED TO A SPATIAL MODEL
OF YEAST MATING POLARIZATION

Ching-Shan Chou
Department of Mathematics
Ohio State University
In systems biology, we would like to...

• Develop a comprehensive model that captures the key features in the modeled system
• Identify the key processes in the system
• Estimate the parameters in the model so that the model has a predictive power
The challenges are...

• With a (relatively) comprehensive model, there are a lot of unknown parameters
• Experimental data are generated in order to estimate the parameters, but
  – Is the system identifiable?
  – Do we have enough data?
  – Are the data useful?
• How to deal with spatial data?
Budding yeast (*Saccharomyces cerevisiae*) as a model system

- Genome fully sequenced (1996)
- Simple geometry
- Pathway understood extensively
- Cdc42 GTPase, the central polarization protein, is highly conserved (human Cdc42 (Cdc42hs) is 80% identical to yeast Cdc42)

Image credit: Masur - Own work. Licensed under Public domain via [Wikimedia Commons](https://commons.wikimedia.org)
Cell cycles of budding yeast
Yeast cell polarization

- Yeast reproduces by budding or mating.
- During mating, cell forms a projection in response to a pheromone signal.
- Polarization occurs via a signaling pathway with positive and negative feedback.
Cell polarization is everywhere

*Cell polarization* is the asymmetric organization of cellular components.

---

**Fundamental to cell differentiation, division, migration, specialized functions, ...**
Yeast mating projection – Shmoo
Accurate mating
$\alpha$-factor induced yeast cell projection

$\alpha$-factor receptor
(Ste2)
α-factor induced yeast cell projection

Cell Cycle Arrest

Cell Polarization

Transcription
All-or-none polarization of proteins

Signal transduction pathway
Signaling transduction pathway

membrane  cytoplasm
Cdc24m

Cdc42a

\[ \frac{\partial [C42a]}{\partial t} = D_{C42a} \nabla^2 [C42a] + k_{42a} [C24m][C42] - k_{42d} [C42a] \]

• **Cdc42** is activated by **Cdc24m**
• **Cdc42a** can deactivate to reform **Cdc42**
• Proteins diffuse along the membrane
Computational domain
Full polarization model

\[ \frac{\partial [R]}{\partial t} = D_{R1} \nabla^2 [R] - k_{RL}[L][R] + k_{RLm}[RL] - k_{Rd}[R] + p_x k_{Rd} \]

\[ \frac{\partial [RL]}{\partial t} = D_{RL} \nabla^2 [RL] + k_{RL}[L][R] - k_{RLm}[RL] - k_{Rd}[RL] \]

\[ \frac{\partial [G]}{\partial t} = D_{G1} \nabla^2 [G] - k_{Ga}[L][G] + k_{Gd}[Gd][Gb] \]

\[ \frac{\partial [Ga]}{\partial t} = D_{Ga} \nabla^2 [Ga] + k_{Ga}[L][Ga] - k_{Ga}[Ga] \]

\[ \frac{\partial [Gb]}{\partial t} = D_{Gb} \nabla^2 [Gb] + k_{Ga}[RL][Gb] + k_{Gd}[Gd][Gb] \]

\[ \frac{\partial [Gd]}{\partial t} = D_{Gd} \nabla^2 [Gd] + k_{Ga}[Ga] - k_{Ga}[Gd][Gb] \]

\[ \frac{\partial [C24m]}{\partial t} = D_{C24m} \nabla^2 [C24m] + k_{24cm}(Gb_{n}^*)[C24c] + k_{24cm}(B1^*)[C24c] \]

\[ \quad - k_{24md}[C24m] - k_{24d}[Cla4a][C24m] \]

\[ \frac{\partial [C42]}{\partial t} = D_{C42} \nabla^2 [C42] - k_{42d}[Gd][C24m] + k_{42d}[C42] \]

\[ \frac{\partial [C42a]}{\partial t} = D_{C42a} \nabla^2 [C42a] + k_{42d}[C42m][C42] - k_{42d}[C42a] \]

\[ \frac{\partial [B1m]}{\partial t} = D_{B1m} \nabla^2 [B1m] + k_{B1cm}[C42a][B1c] - k_{B1md}[B1m] \]

\[ \frac{\partial [Cla4a]}{\partial t} = k_{Cla4a}[C42a^*] - k_{Cla4a}[Cla4a] \].

- Large nonlinear PDE – expensive to solve!
- 35 unknown parameters
Full polarization model

Quantity of Interest: Extent of polarization of Cdc42a at steady state (how concentrated is active Cdc42?)

\[ PF(C42a) = 1 - 2 \frac{S_p(C42a)}{SA} \]

How much Cdc42 is activated?

\[ z = PF(C42a) \times \frac{(ax)^n}{1 + (ax)^n} \]

\[ a = \frac{2*SA}{C42_t} \quad \text{and} \quad x = \max(C42a) \]
Introduction

Model $f(x, \theta)$

Data

How do we minimize computational cost?

How do parameters affect model output?

Can data inform choice of parameters?
Parameter sensitivity

**Question:** Can we quantify the effects of parameters on the behavior of the system?

Let $z$ be a scalar quantity of interest, $z = f(\theta_1, \theta_2, \ldots, \theta_N)$

**Local sensitivity:** $S_j = \frac{\partial z}{\partial \theta_j}$ at some point $\theta^*$

- Can be computed analytically or approximated by finite difference
- Only gives information at one point – what if we don’t know the parameter values?
Parameter sensitivity

Correlations
- Pearson correlation coefficient
- Rank/Spearman correlation coefficient
- PRCC (accounts for effects of other parameters)

Variance-based sensitivity measures
- Variance decomposition
  \[ \text{Var}(z) = \sum_{i=1}^{N} V_i + \sum_{i<j} V_{ij} + \ldots + V_{1,2,\ldots,N} \]
- Main effect index \( \frac{V_j}{\text{Var}(z)} \), Total-effect index

Derivative-based sensitivity measures
\[ E\left[ \frac{\partial z}{\partial \theta_j} \right], \ E\left[ \left| \frac{\partial z}{\partial \theta_j} \right| \right], \ E\left[ (\frac{\partial z}{\partial \theta_j})^2 \right] \]
Methods: Parameter estimation

**Question**: Given experimental data, can we estimate the parameters?

- What are the most likely parameter values?
- How much uncertainty is there in our parameter estimates?

Bayesian inference allows us to recover a probability distribution of the parameters.
Methods: Parameter estimation

Markov chain Monte Carlo (Metropolis-Hastings):

1. Choose initial parameter set $\theta_0$.
2. For $n = 0, 1, 2, \ldots$
   i. Sample a new parameter set $\theta^*$ close to $\theta_n$
   ii. Calculate the acceptance ratio $\alpha = \frac{L(\theta^*)}{L(\theta_n)}$.
   iii. Accept or reject
      • Generate a uniform random number $u$ on $[0,1]$
      • If $u \leq \alpha$, set $\theta_{n+1} = \theta^*$ (accept).
      • If $u > \alpha$, set $\theta_{n+1} = \theta_n$ (reject).
Methods: Polynomial surrogate

How do we sample a high-dimensional parameter space when each sample requires an expensive model evaluation (e.g. a large PDE)?

Short answer: We don’t.

**Solution**

We can fit a polynomial to $z$ in terms of the parameters $\theta_1, \theta_2, ..., \theta_N$.

$$z \approx P(\theta_1, \theta_2, ..., \theta_N)$$

Polynomial basis depends on probability distribution of the parameters.

The polynomial serves as a surrogate for the full model.
Generalized Polynomial Chaos (gPC)

Suppose $\theta_i \sim U[-1,1]$ for all $i = 1, \ldots, N$.

For $i = 1, \ldots, N$, let $\{\phi_m(\theta_i)\}_{m=0}^{d}$ be a set of orthogonal polynomials, i.e.

$$\frac{1}{2} \int_{-1}^{1} \phi_m(\theta_i) \phi_n(\theta_i) dy_i = h_m^2 \delta_{mn}$$

For the uniform distribution, this gives the Legendre polynomials.

These form a basis for univariate polynomials up to degree $d$. 
Generalized Polynomial Chaos (gPC)

The corresponding N-variate orthogonal polynomial space is defined to be

\[ W_N^P = \bigotimes |d| \leq P \left( \operatorname{span} \{ \phi_m(\theta_i) \}_{m=0}^{d_i} \right) \]

where \( d = (d_1, d_2, \ldots, d_N) \) is a multi-index with \( |d| = \sum_{i=1}^{N} d_i \).

The \( P \)th order gPC approximation of \( z \) can be obtained by projecting \( z \) onto the space \( W_N^P \):

\[ z_N^P(\theta) = \sum_{m=1}^{M} \hat{\theta}_m \Phi_m(\theta) \]

**Options:** Orthogonal projection (Galerkin), polynomial fitting
Polynomial fitting

\[ z_N^p(\theta) = \sum_{m=1}^{M} \hat{z}_m \Phi_m(\theta) \]

Want to solve for polynomial coefficients, given samples \( \theta^{(1)}, \ldots, \theta^{(K)} \).

Let \( x = (\hat{z}_1, \ldots, \hat{z}_M) \) be the vector of coefficients.

\[ Ax = b \]

\( A \) is a matrix where each row corresponds to one sample and each column corresponds to one basis polynomial.
\( b \) is a vector of model output at the sample points.
Dimension of polynomial space is $\binom{N+P}{N}$, where $P$ is the degree.

### Polynomial fitting

<table>
<thead>
<tr>
<th># of parameters</th>
<th>Polynomial degree</th>
<th># of coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>252</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>3003</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>53130</td>
</tr>
</tbody>
</table>

### Polynomial fitting:

<table>
<thead>
<tr>
<th>Undersampling</th>
<th>Exact fitting</th>
<th>Oversampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed sensing</td>
<td>Interpolation</td>
<td>Least squares approximation</td>
</tr>
</tbody>
</table>
Least squares approximation

- # of samples > # of coefficients
- Generally no exact solution
- Find polynomial that minimizes the squared error,
  \[ \sum_{i=1}^{n} (P(\theta_i) - b_i)^2 \]
- Equivalent to solving \( A^T A x = A^T b \)
Compressed sensing (in a nutshell)

- # of samples < # of coefficients
- Infinitely many solutions
- Find solution with minimal $l_1$-norm (the “sparsest” polynomial)
Methods: polynomial surrogate

Now that we have a polynomial surrogate,

- \( S_j = \mathbb{E} \left[ \frac{\partial z}{\partial \theta_j} \right] \) can be **analytically computed** – no sampling required.
- Markov chain Monte Carlo (MCMC) samples require only a polynomial evaluation.

The only model evaluations required are to fit the polynomial.
Methods: summary

1. Determine quantity of interest
2. Perform sensitivity analysis
3. Can parameter count be reduced?
   - Yes: Fix non-sensitive parameters
   - No: Parameter estimation
Results: Toy model

Toy model: ODE model of heterotrimeric G-protein cycle

\[
\begin{align*}
\frac{d[R]}{dt} &= -k_{RL} L [R] + k_{RLm} [RL] - k_{Rd0} [R] + k_{Rs} \\
\frac{d[RL]}{dt} &= k_{RL} L [R] - k_{RLm} [RL] - k_{Rd1} [RL] \\
\frac{d[G]}{dt} &= -k_{Ga} RL [G] + k_{G1} [Gd][Gb] \\
\frac{d[Ga]}{dt} &= k_{Ga} RL [G] - k_{Gd} Ga
\end{align*}
\]

- ODE – easy to solve.
- 8 parameters: 6 have been measured in experiments, 2 have been estimated via optimization*.

### Toy model – Data

**Data output:** Fraction of free \( G\beta\gamma \) at various time and \( \alpha \)-factor

<table>
<thead>
<tr>
<th>( \alpha )-factor (( L ))</th>
<th>Time (( T ))</th>
<th>( \alpha )-factor (( L ))</th>
<th>Time (( T ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 nM</td>
<td>10 s</td>
<td>1 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>30 s</td>
<td>2 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>60 s</td>
<td>5 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>120 s</td>
<td>10 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>210 s</td>
<td>20 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>300 s</td>
<td>50 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>450 s</td>
<td>100 nM</td>
<td>60 s</td>
</tr>
<tr>
<td>1000 nM</td>
<td>600 s</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Toy model – Fitting with 2 parameters

- 5th order polynomial errors
- Error of least squares polynomial, $n = 1000$
- Cost to compute polynomials
- Cost to compute polynomials
Toy model—Results

Mean sensitivity

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{Ga}$</td>
<td>0.42497</td>
</tr>
<tr>
<td>$k_{Gd}$</td>
<td>-0.32092</td>
</tr>
</tbody>
</table>

Estimated
MSE = 4.1 x $10^{-4}$

Optimal
MSE = 1.3 x $10^{-4}$

Time-course

Dose-response

Toy model – Approximating all 8 parameters

Approximating all 8 parameters:

• $5^{th}$ degree polynomial with 1500 samples
• Error: Mean 0.00221, Std 0.047492
• Data: Fraction of free $G\beta\gamma$ at various time points, $\alpha$-factor levels

\[
\frac{d[R]}{dt} = -k_{RL}[L][R] + k_{RLm}[RL] - k_{Rd0}[R] + k_{Rs}
\]
\[
\frac{d[RL]}{dt} = k_{RL}[L][R] - k_{RLm}[RL] - k_{Rd1}[RL]
\]
\[
\frac{d[G]}{dt} = -k_{Ga}[RL][G] + k_{G1}[Gd][Gbg]
\]
\[
\frac{d[Ga]}{dt} = k_{Ga}[RL][G] - k_{Gd}[Ga]
\]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{RL}$</td>
<td>0.081921</td>
</tr>
<tr>
<td>$k_{RLm}$</td>
<td>-0.032349</td>
</tr>
<tr>
<td>$k_{Rs}$</td>
<td>0.0092325</td>
</tr>
<tr>
<td>$k_{Rd0}$</td>
<td>0.001139</td>
</tr>
<tr>
<td>$k_{Rd1}$</td>
<td>-0.062337</td>
</tr>
<tr>
<td>$k_{G1}$</td>
<td>0.00055961</td>
</tr>
<tr>
<td>$k_{Ga}$</td>
<td>0.30979</td>
</tr>
<tr>
<td>$k_{Gd}$</td>
<td>-0.26078</td>
</tr>
</tbody>
</table>
Toy model – 8 parameter MCMC results

Estimated
MSE = 6.4 x 10^{-4}

Optimal
MSE = 1.3 x 10^{-4}
Findings

• System is most sensitive to activation and deactivation of Gα

• Polynomial surrogates produce similar results to optimization method

• Sensitive parameters have lower uncertainty in estimation
  – High uncertainty occurs in parameters that have been experimentally measured
Large nonlinear PDE – expensive to solve!

35 unknown parameters
Full polarization model – Polynomial fitting

- 5th order polynomial $\Rightarrow$ 658,008 coefficients
  $\rightarrow$ Undersample and use $l_1$-minimization

- With 5000 samples, mean absolute error was 0.12.
Most parameters have small sensitivity.

→ Reduce parameter count!

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
<th>Parameter</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{B1m}$</td>
<td>-0.000240521</td>
<td>$C42_t$</td>
<td>0.00831949</td>
</tr>
<tr>
<td>$G_t$</td>
<td>-0.00025359</td>
<td>$D_{Gb}$</td>
<td>-0.0097571</td>
</tr>
<tr>
<td>$D_{Ga}$</td>
<td>-0.000704648</td>
<td>$B1_t$</td>
<td>0.0172396</td>
</tr>
<tr>
<td>$D_{RL}$</td>
<td>-0.000734672</td>
<td>$k_{24d}$</td>
<td>-0.0184795</td>
</tr>
<tr>
<td>$D_{Gd}$</td>
<td>-0.00106669</td>
<td>$C24_t$</td>
<td>0.0203875</td>
</tr>
<tr>
<td>$k_{Ga}$</td>
<td>0.00151801</td>
<td>$k_{Cl4a4a}$</td>
<td>-0.0210338</td>
</tr>
<tr>
<td>$k_{Rs}$</td>
<td>0.0015794</td>
<td>$k_{Cl4d}$</td>
<td>0.0247716</td>
</tr>
<tr>
<td>$R_t$</td>
<td>-0.00167628</td>
<td>$k_{B1cm}$</td>
<td>0.0463723</td>
</tr>
<tr>
<td>$k_{Rd0}$</td>
<td>0.00232993</td>
<td>$k_{24cm0}$</td>
<td>-0.0521665</td>
</tr>
<tr>
<td>$D_{c24m}$</td>
<td>-0.00233107</td>
<td>$k_{B1mc}$</td>
<td>-0.0550074</td>
</tr>
<tr>
<td>$k_{G1}$</td>
<td>0.00234184</td>
<td>$D_{c42a}$</td>
<td>-0.0557088</td>
</tr>
<tr>
<td>$k_{24mc}$</td>
<td>-0.00243174</td>
<td>$h$</td>
<td>0.056832</td>
</tr>
<tr>
<td>$k_{RLm}$</td>
<td>0.00348159</td>
<td>$D_{c42}$</td>
<td>0.0624607</td>
</tr>
<tr>
<td>$k_{Rd1}$</td>
<td>0.00401553</td>
<td>$q$</td>
<td>0.0744311</td>
</tr>
<tr>
<td>$k_{Gd}$</td>
<td>0.0040475</td>
<td>$k_{42d}$</td>
<td>-0.0962736</td>
</tr>
<tr>
<td>$D_{R}$</td>
<td>-0.00517784</td>
<td>$k_{24cm1}$</td>
<td>0.0999564</td>
</tr>
<tr>
<td>$D_{G}$</td>
<td>0.00608915</td>
<td>$k_{42a}$</td>
<td>0.110069</td>
</tr>
<tr>
<td>$k_{RL}$</td>
<td>-0.00789241</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Full polarization model – 15 parameters

Fit new polynomial in reduced space with 6000 samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{24d}$</td>
<td>0.00137156</td>
</tr>
<tr>
<td>$k_{Cla4a}$</td>
<td>-0.00400381</td>
</tr>
<tr>
<td>$k_{24cm0}$</td>
<td>-0.0131764</td>
</tr>
<tr>
<td>$B_{1t}$</td>
<td>0.0174327</td>
</tr>
<tr>
<td>$k_{Cla4d}$</td>
<td>0.0247462</td>
</tr>
<tr>
<td>$C_{24t}$</td>
<td>0.043892</td>
</tr>
<tr>
<td>$k_{B1cm}$</td>
<td>0.0553076</td>
</tr>
<tr>
<td>$k_{B1mc}$</td>
<td>-0.0557505</td>
</tr>
<tr>
<td>$D_{c42a}$</td>
<td>-0.059246</td>
</tr>
<tr>
<td>$q$</td>
<td>0.0607483</td>
</tr>
<tr>
<td>$D_{c42}$</td>
<td>0.0695064</td>
</tr>
<tr>
<td>$h$</td>
<td>0.0774857</td>
</tr>
<tr>
<td>$k_{24cm1}$</td>
<td>0.0977702</td>
</tr>
<tr>
<td>$k_{42d}$</td>
<td>-0.127457</td>
</tr>
<tr>
<td>$k_{42a}$</td>
<td>0.137561</td>
</tr>
</tbody>
</table>
Polynomial surrogates can also be applied to optimization methods, e.g. simulated annealing (SA).

- SA produces better fit.
- MCMC shows that we must be careful with interpretation.
  - ‘optimal’ parameter sets are not unique.
Summary

What we’ve done:

1. Fit polynomial surrogate using 5000 samples
2. Parameter sensitivity analysis to reduce parameter count
3. Fit new polynomial in reduced space with 6000 samples
4. Parameter estimation via MCMC (chain length 2,000,000)

Cost of estimation: 11,000 model evaluations vs. 2,000,000

→ 180-fold reduction in cost
→ Parallelizable
Conclusions

• System is most sensitive to parameters associated with Cdc42 cycle and feedback loops.

• Current data is not enough to estimate parameters, but indicates difference between active and inactive Cdc42.

• Can use time series data, other protein species.
  – Problems: time delay, correlation between species.

• Addition of other protein species – Model selection.